

Vocal Assistants for Reading News

Jennifer Swaminathan
University of Neuchatel
Neuchatel, Switzerland

Email: jennifer.swaminathan@unine.ch

Leonardo Angelini
University of Fribourg
Fribourg, Switzerland

Email: leonardo.angelini@hefr.ch

Abstract—Vocal assistants have become popular in recent years. They are virtual assistants, that can be activated by providing voice commands. They make daily tasks more efficient and easier. These assistants are also being used for providing information on the news. This paper reviews recent research on conversational agents for the purpose of reading news. We answer our research questions by finding techniques that are used in the research papers to retain users’ interaction, for example using SSML techniques to make changes to the monotonous tone of the voice assistants, which improves users’ experience and interaction with the vocal assistant. We compare the different techniques and synthesize them. In the end, we provide the components needed to design a vocal assistant for reading news and highlight how there is increased concern for privacy.

I. INTRODUCTION

Conversational agents are software programs that interact with humans in natural language. With a conversational user interface(CUI), users can interact in real-time. These agents can be text-based chatbots, social chatbots, or voice assistants like Alexa, Siri, Google hub.

Voice assistants are present in many houses, mobile phones and, cars. They make simple tasks of daily life seem easy using simple language. They remove the need for user interfaces like keyboards and screens. For example, voice commands can be used to place timers without the use of an input device or to check for the weather. These voice assistants can be connected to smart appliances to control these appliances. This makes having these interfaces efficient.

These voice assistants are helpful for individuals with disabilities. They provide them with information by reading them out loud.

Vocal assistants can be proactive, by reminding users of their schedules, and to take medicine. But these proactive smart speakers can be disruptive by providing vocal notifications, during unintended times. Cha et al. in [2] suggest, context-aware speakers or proactive interactions so that the users are not disrupted from their routine.

Some vocal assistants consist of a screen, also known as a smart display, which can provide visual information along with synthesized voice. Users of these devices are required to interact with the agent by voice commands.

Vocal assistants can be useful for consuming news through a digital format. The use of traditional newspapers has been declining in recent years, as more people turn to online sources for their news. Some vocal assistants offer a feature where they can read the headlines or present news in a monotone manner,

making it convenient for users to stay informed without having to actively seek out the information themselves. For example, instead of searching for news on smartphones, vocal assistants can read the news to the users.

Vocal assistants for reading news can increase the engagement of the user through interactive and informal conversation. Also, using Natural Language Processing and Text-to-Speech, news can be translated into other languages and presented in a natural way, increasing the accessibility of the news for non-native speakers as seen in research done by Carvalho et al in [1]. Additionally, vocal assistants can serve as a way for users to discover other news sources and broaden their knowledge, rather than relying on a single source.

However, there are limitations to vocal assistants. When users encounter problems with the devices, they may lose trust in them as discussed by Dambanemuya et al in [6]. Additionally, there are concerns about privacy, where the assistant is gaining information about the users, Rao et al. discuss about this concern in [17]. Additionally, some people may be uncomfortable having a device that can potentially record conversations in their homes without their knowledge, leading to privacy concerns which are shown by Lau et al. in [13]. It’s important for users to be aware of the privacy settings on the device and how their data is being collected and used. Another limitation is that vocal assistants can only provide one search result at a time, unlike search engines which can provide multiple results. Therefore, it’s important for the results provided by the vocal assistants to be accurate and relevant, to maintain user trust and satisfaction as suggested by Paranell et al in [16].

Designing a vocal assistant to maintain interaction is crucial for its success. One technique used to retain this interaction is provided by Sahijwani et al. in [18]. They propose a voice assistant model where the agent quickly changes the topic when the user is no more interested in a topic as shown in Fig 1. We can see how the chatbot provides ideas for discussion to retain users interaction.

This literature review synthesizes techniques to maintain the engagement of the users in recent research papers. We answer our research questions

Q1: What are the methods used to retain user interaction?

Q2: Are there improvements in user experience with these approaches?

The structure of this paper is as follows. Section 2 presents definitions of terms, namely, conversational agents(CA), voice

Bot : Hey, did you see the recent movie Joker?
 User: Yes, I loved it. It's pretty awesome.
 Bot : That's nice. I'm guessing you like superhero movies?
 User: Not really, but Joker is just on another level. Loved the character and the aesthetic.
 Bot : I see. Would you like to talk about other movies like the Joker?
 User: I don't want to talk about movies.
 Bot : There is a trending technology story about Apple's new AirPods Pro. Want to hear it?
 User: Yes.
 Bot : Wired on Oct 28th reported: Apple revealed a new version of AirPods on its website today, adding the totally wire-free headphones to its list of products. I have more news on this topic. Would you like to hear?

Fig. 1. Conversation with chatbot extracted from [18]

assistants, podcasts, and the key components needed to build these voice assistants. Section 3 presents the methodology followed in our analysis. Section 4 provides the results obtained from our review. Section 5 provides a synthesis and discussion of our review and provides the components needed to design a vocal assistant for reading news. Section 6 provides the conclusion and possible future work.

II. DEFINITIONS

In this section, we define terms used in designing Conversation Agents. Conversational agents or chatbots are programs that simulate conversations like human users. Text chatbots are used for customer service during online shopping, or messenger chatbots on social media. Voice chatbots or voice assistant, interact with users via speech. Example of voice assistants are Siri, Google, and Alexa.

A common use for conversation with these chatbots is for information retrieval(IR). These agents, text chatbots, and voice assistants have similar algorithms. The difference between the two is the way the two receive and process the information. Textual chatbots use Natural Language Processing and Machine learning techniques to understand and respond to the user. Voice assistants, on the other hand, use speech recognition technology to understand user input in the form of speech and Text-to-Speech (TTS) technology to respond to the user in speech.

Parnell et al. in [16], performed evaluations to find that vocal assistants are preferred to textual chatbots, mainly because vocal assistants help in task completion along with providing information.

Fig 2, shows the architecture of how a voice assistant works in Alexa [6]. Voice assistants can perform tasks like IR, delivering news, and controlling smart home appliances. These voice assistants are used by first starting with a 'wake up' word like 'Alexa, Siri, Hey Google', followed by the request for information or a voice command to control the smart appliances. This voice query is then transcribed using speech recognition, machine learning, and Natural language understanding(NLU) algorithms. From these algorithms, the intent of the query is obtained using sentiment analysis, and entity recognition. Using IR related documents are obtained from the internet. The result is converted into a textual



Fig. 2. Architecture of voice assistant extracted from [6]

response which is converted to audio files through the TTS algorithm and provided to the users through smart speakers.

To shorten the information obtained from the internet, text summarization techniques are suggested by Gupta et al. in [8]. Automatic summarization eliminates the need of reading long texts. There are two types of summarization, extractive and abstractive. The extractive technique identifies the relevant information from the original document and extracts only those texts. Abstractive summarization generates new sentences and phrases by capturing the meaning of the original relevant documents.

Abstractive summarization is preferred to extractive as it generates own sentences. Abstractive summarization is trained on large datasets to learn the meaning of the text, context, and word relationship. The authors propose using pre-trained language models with fine-tuned transformers to obtain sound and fluent summaries for a document. These pre-trained language models are used in voice assistants to provide information to the users. For example, Laban et al. in [12], uses PEGASUS model, for summarization.

Question Answering Generator(QAGen) are models that retrieve relevant information from the internet and generate answers using summarization techniques which are used by Liu et al. in [15].

Question Generation(QGen) models are used, to formulate follow-up questions, that the users might be interested to know. This helps to keep the flow of the conversation. QGen can also be used to generate Quiz to interact with the users as shown in [14]. QAGen and QGen helps in maintaining the naturalness of information provided by the voice assistant.

Voice assistants can provide news using podcasts. Podcasts are digital audio files that are available on the internet, containing different contents. The main purpose is to, entertain its audience and provide information on current affairs. These are created by individuals or groups. But instead of depending on humans, podcasts can be automated. Automated podcasts differs from summarization techniques used in vocal assistants. Automated podcasts are initially manually created and presented to the user using TTS by translating to other languages as shown by Carvalho et al in [1]. Summarization techniques can be incorporated into the podcast when the user asks questions related to the podcasts which is used by Laban et al in [12].

III. METHODOLOGY

In our literature review on Vocal assistants for reading news, we started our research by searching Semantic Scholar and

ACM digital libraries. Following were the queries we used "Vocal Assistants news", "Conversational agents news", "news voice assistant question answer generation" and "summarization news", and applied filters to restrict research papers in the past 5 years. We also looked for papers related to smart displays to see if screens can be advantageous for reading news. From the results, we chose papers related to our research questions and chose 17 papers.

IV. RESULTS

In this section, we summarize our findings on vocal assistants into the following categories: Question-Answer generation, proactive, personalized, automated podcasts, and modulated voice assistants.

A. Question-Answer generation in vocal assistants

Traditional news briefings are monotonous and one-way communication. To maintain user engagement while providing news through voice assistants interaction between the user and voice assistant is key. In order to achieve interactivity, Jung et al in [10] describe the process of creating a corpus of information from the news topics and generating expected Q&A. These are given to TTS to produce audio files to give the users when a query is asked. This allows interactivity between the user and the voice assistants. To verify if users preferred interaction during a news briefing, the authors performed user studies. The participants preferred the interactions but were not comfortable asking questions or interrupting the agent during the briefing. Additionally, when given the opportunity to ask questions, the participants often forget what they wanted to ask.

However, when there are interactions between voice assistants and users, the assistant could provide the same response using different phrases. In order to combat the issue of repetitive response, Laban et al in [11] propose a new design. They collect content from multiple news sources, split the corpus into paragraphs, and use Paragraph / Question bipartite graph (P/Q graph) to keep track of the conversation, as shown in Fig 3. Here, when a user asks a question, paragraphs with the answers are mapped. The system prioritizes paragraphs that answer the most questions and avoids using paragraphs that do not answer the questions. When a paragraph is used to provide the answer, the paragraph is not revisited to answer the users question.

The authors performed an evaluation on this model and found that participants had longer conversations with the device. The drawback of this model was that the model was unable to handle small talk and some QA generated were inaccurate.

Another method to formulate Q&A for the voice assistant, Liu et al in [15], propose a QAGen model using Clue, Style and Answer aware parameters. This model is used to imitate the way humans ask questions and allows human-like conversations with the voice assistant. Their idea is that given a paragraph there are different types of questions that can be formulated, as shown in Fig 4. Using the additional parameter

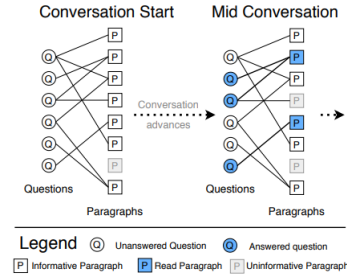


Fig. 3. Avoid repetition in Chatbots extracted from [11]

The fight scene finale between Sharon and the character played by Ali Larter, from the movie Obsessed, won the 2010 MTV Movie Award for Best Fight.
Answer: MTV Movie Award for Best Fight Clue: from the movie Obsessed Style: Which Q: A fight scene from the movie, Obsessed, won which award?
Answer: MTV Movie Award for Best Fight Clue: The fight scene finale between Sharon and the character played by Ali Larter Style: Which Q: Which award did the fight scene between Sharon and the role of Ali Larter win?
Answer: Obsessed Clue: won the 2010 MTV Movie Award for Best Fight Style: What Q: What is the name of the movie that won the 2010 MTV Movie Award for Best Fight?

Fig. 4. HumanWay questions extracted from [15]

of Answer, Style, and clue questions are generated in a human-like style. They say that Answer, style, and clue are dependent parameters. For example, for the question "A fight scene from the movie Obsessed won which award". "Clue" is the information from the input query 'from the movie Obsessed'. The Style is 'Which', the authors propose to use 9 styles, like, 'yes-no', 'when', 'how' and more. From these, they obtain the 'Answer' from the paragraph to produce the QA generator with ACS parameters.

By providing all the news information, how do we measure if the users have retained the information provided? Lelkes et al in [14], propose an automatically generated quiz-style question generation to check the attention level of the users after news stories. In order to model this they use two components, Question Answer Generator(QAG) and a distractor generator (DG). QAGen obtains the question and answers from the narrated news story and DG is a tool to provide plausible but incorrect answers. The authors evaluated their model of Quiz Style QA generator by performing a survey on Google survey site with quizzes for 2 months and found participants found the automatic multiple choice Quiz enjoyable and educational.

B. Proactive vocal assistants

Proactive vocal assistants can increase news consumption, by providing news briefings to users at appropriate moments. But these proactive assistants can not provide messages when the user is not at home or when they are occupied. Cha et al. in [2], propose a method to design a proactive vocal assistant, to find the interruptible or opportune moments at home for a

Type	Description	Examples
Generic	Proposing to talk about news without a specific entity/headline.	"There is so much going on in the world. Would you like to talk about the news?"
Trending News	Proposing to tell news on a trending topic from Bing News/Google Trends.	"How about a recent sports story titled 'Astros solidify comeback'? Want to hear it?" "There is a trending technology story about Apple AirPods Pro. Want to hear it?"
News Briefing	Proposing to provide a 45-second summary of the news from Reuters.	"I prepared a 45-second overview of the news for you. Want to hear it?"
Entity based (Current Conversation)	Proposing to tell news about the user's favorite sports team (if they mentioned it earlier), or any entity that the user asked a question about. (How old is the Dalai Lama?)	"Do you want to hear the latest sports updates on your favorite team?"
Entity based (Past Conversation)	Proposing to tell news about an entity mentioned by the user in past conversations.	"Last time we spoke, you mentioned Imagine Dragons. Would you like to hear the latest news on Imagine Dragons?"

Fig. 5. News recommendation extracted from [18]

conversation. To measure these moments, sensors are required to make observations. They performed user study to determine the appropriate times for a proactive vocal assistant to initiate a conversation with users. Through a week-long experiment with participants, the authors found that domestic routines and spatial relationships should be considered before a proactive speaker starts a conversation. They also found that participants were generally willing to be interrupted when returning from outside, but not during tasks requiring concentration, such as studying or meetings. The study highlights the benefits of proactive speakers in providing information but also acknowledges the challenges related to privacy and disruption. The authors suggest the use of context awareness to mitigate these challenges.

Wei et al in [19], find methods on how to interrupt the users when providing proactive messages. To start a conversation, should the assistant just start talking or ask for permission, or use music to interrupt the user? These methods need to be implemented to retain the user's interaction and not be disruptive. A week-long field test was conducted with a proactive speaker in participants' homes to evaluate the model. Three different methods of interrupting the users were used: 1) an utterance of "hey are you available", 2) an earcone music sound similar to smartphone notifications, and 3) a baseline with no starter sentence. The evaluation results showed that participants preferred the utterance mode as it felt more human-like and was easier to converse with. Some participants found the proactive addition to be helpful and could multi-task while conversing with the device. It was also found that people preferred conversing when they were closer to the device.

C. Personalised vocal assistant

To provide personalized recommendations to the user to maintain interaction with the user and voice assistants, Sahin-jani et al in [18], analyzed the logs of historical conversations that were had with the voice assistant and found the users generally asked questions based on different categories, for example, news from a particular category, news about a specific entity, news briefing, and generic. Based on this they recommend news on 5 types, Generic, Trending News, News Briefing, Entity based (current conversation or past conversation) as shown in Fig 5. They performed evaluations on real-time Alexa users. The results showed that there were higher acceptance rates for entity-based news recommendations using

Topic	minimal	keyword	full sentence
Weather	"sunny, 5 to -3 degrees"	"Weather tomorrow: sunny, 5 to -3 degrees"	"The weather tomorrow is gonna be sunny with a high of 5 degrees and a low of -3 degrees."
Calendar	"Lunch with Steven, 12PM. Dinner with Ann, 7PM."	"Meetings tomorrow: Lunch with Steven, 12PM. Dinner with Ann, 7PM."	"You have two appointments. At 12 PM there is Lunch with Steven and at 7 PM dinner with Ann."
News	"BBC - Australia not intimidated by Facebook news ban"	"News today: BBC - Australia not intimidated by Facebook news ban"	"Here is what I found. BBC's latest headline is: Australia not intimidated by Facebook news ban."
General knowledge	"328 million."	"US population: 328 million."	"The total size of the US population amounts to 328 million."

Fig. 6. Styles of responses in VA extracted from [9]

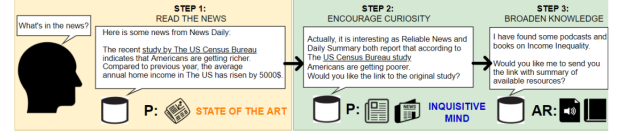


Fig. 7. Inquisitive mind extracted from [7]

historical conversation. Thus providing a personalized recommendation.

Another method to personalize the recommendation is by providing responses based on the preference of the user. Haas et al in [9], perform a study on how the length of the response can influence the attention of the user. Fig 6, shows how different topics can produce different styles of answers, namely, minimal, keyword, and full sentence. Based on the preference of the user, different formats of response can be provided to the user. They performed user study using these three styles. The results showed that minimal response was efficient and the younger generation preferred them. However, only 13 % wanted shorter answers for News readings.

Contrary to a preference for personalized recommendation, Chen et al in [4], performed a semi-structured interview to find users preferred personalized as well as non-personalized recommendations for diversity. This is due to a lack of diversity when a voice assistant provides only personalized recommendation, causing a Filter Bubble.

In order to combat the Filter Bubble effect, Dubiel et al. in [7], suggest encouraging the curiosity of the users by providing contradicting facts from other sources using contrastive sentiment analysis. Also, to broaden the knowledge of the users, they suggest providing further readings and podcasts as shown in Fig 7. This can lead to more news consumption.

D. Automated podcasts and modulated voice assistants

Vocal assistants make use of TTS algorithms to convert the generated text into speech. This TTS has several features and can be used for various purposes. In this subsection, we list some of their uses.

Firstly, TTS as a translator. Podcasts have become popular in recent years. Carvalho et al. in [1] suggest, using TTS algorithms to create news podcasts automatically in Portuguese by obtaining information from news websites. They translate the original manuscript to podcasts and reduce the need for human intervention. The authors performed user studies using 3 voices, human, Synthesized voice, and hybrid for reading news. The evaluation showed acceptance of the integration of

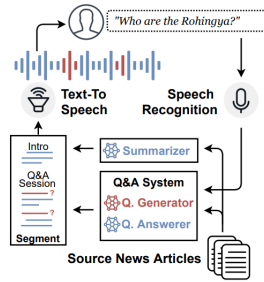


Fig. 8. NewsPod Flowchart extracted from [11]

voices and found hybrid voices were acceptable. This methodology increases news consumption and provides accessibility to the news.

Next, we talk about the advantage of TTS to generate podcasts with different voices. Laban et al in [12], curate automatically generated podcasts for news (NewsPod) on several topics. The podcast is divided into segments and each segment is on a different topic in Question and Answer format. To retain the attention of the users, a key design feature is they use distinct voices for questions and responders to simulate human conversation. This NewsPod allows users to interact with them and the questions are automatically answered. The flowchart used in this is shown in Fig 8. A segment is created on a news topic. They make use of a Summarizer and Q&A system, which automatically generates Questions and Answers. This is later converted from Text-To-Speech. The user can interrupt the conversation and ask a question. Using speech recognition, the question and answer are generated from the source. The authors performed user studies to test their model. They found that users preferred modulated sessions to monotonous article. They found it natural. They preferred when the podcasts had a break before moving to the next topic so that the participants could ask questions. The limitations of this model were the participants were not satisfied with the questions they asked. They found the responses confusing and irrelevant.

Another feature of TTS is modulation. Chan et al in [3], propose using synthesized voice of friends and family to retain the engagement of the users. This feature can bring a sense of closeness to physical distance. This can be done by recording the voice of a friend or family member and then using TTS technology to synthesize a voice that sounds similar to the person's voice. They performed surveys on participants by performing conversation with 1) synthesized KinVoice and 2) generic voice. They used conversation like reminders, to play music. The results showed that the Kin voices were engaging for daily tasks, but others found it different from the original voice.

Another use of TTS in voice assistants is highlighting the key answer by modifying the voice. Chuklin et al. in [5], propose using prosody modification such as pauses, changes in speaking rate, and pitch, to highlight key points in audio responses from voice assistants. They use Speech Synthesis Markup Language(SSML) and prosody parameters to achieve

these modifications, which are crucial for providing naturalness in synthetic speech and highlighting key answers in voice assistants without screens. Through experiments conducted to evaluate the model, the authors found that by lowering the speaking rate and increasing pitch, answers could be more easily identified. However, they also found that these modifications resulted in unnatural speech. The study highlights the potential benefits of prosody modification in improving the effectiveness of voice assistants but also notes the trade-off with naturalness of speech.

V. DISCUSSION

From the results of our literature review, given in Section IV, we present a synthesis table of the models studied in TABLE I. This synthesis compares the aim of the paper, the methodologies used to retain the attention of the user, the advantages, disadvantages of the methods, and the evaluations performed. The table also indicates if the studied papers used vocal interface, related to news, and interactivity with the user.

To answer our Q1, methods used to retain user interaction, we list the techniques studied. TTS algorithm is used to increase the interaction. They are used in different methods like translation, prosody modification, and synthesized voice generation. QA generators are used to summarize the news instead of reading lengthy information. QAGen are also used to generate quizzes to entertain and educate the users. Interaction with the voice assistants helps in engagement of the user. Personalized recommendations using preferences of the users can also increase the consumption of news.

To answer our Q2, are there any improvements in user experience with the addition of the techniques, we summarize our findings. Improved QAGen can help in avoiding repetition of the responses so that users do not lose trust in these devices. Advances in TTS technology have lead to changes in the voices so that synthesized voices can sound like family and friends. This can help bridge physical distances. Proactive assistants can provide news headlines by being aware of the context of the user. Finally, using data collected from the user we can provide personalised recommendations to the user, by either providing information using short response or full descriptions.

Furthermore, as people increase their trust in voice assistants, they could provide content without verifying their source. It is possible to have fake and unreliable news. In order to avoid fake news delivery by the voice assistants, its important that they cite their source. People reveal their ideologies while conversing with the assistant, raising ethical and privacy concerns. For example, Rao et al in [17], provide guidelines for delivering information in an ethical manner, with a focus on ensuring that only fact-checked news and information are provided to users. Additionally, the article suggests that user privacy should be taken into account and that users should be made aware of the level of self-disclosure required for interaction. There are also errors due to speech recognition. For example, voice assistants tend to answer questions while users do not intend to use them.

TABLE I
SYNTHESIS

	Newspod [12]	Inquisitive Mind [7]	Human Way Questions [15]	Prosody Modification [5]	Quiz style-news [14]	Tell me more [10]	Whats the latest (Question News)[11]	TTS Technology in News Podcast[1]	Would you like to hear news [18]	Keep short [9]	it Hello there, is now a good time[2]	Kin Voice [3]	Understand Proactive smart speakers[19]
Year	2022	2019	2020	2019	2020	2019	2021	2022	2020	2022	2020	2021	2021
Aim	Automated news podcasts	Encourage curiosity: sending links and articles	Q&A for a dataset: with varying questions like humans	Evaluating the usefulness of audio transformation	Retain quiz info with format interaction	Design a CA with interaction	Chatbot to answer any type of questions based on news	Automatic news podcasts in different language	Maintain engagement	VA answer with short answers	Proactive smart speaker	How users perceive the kin voice	Note opportune moments in proactive speakers
How? Re-attention	SSML- for different voices	Increase curiosity by providing additional info	Provides diverse question and answers	SSML for speed, rate, pitch	Check user is attentive with Quiz	by interacting	Avoid repetition, tracking conversation	Using TTS to add different voices	Personalised recommendation	using short answers	Proactive speaker with minimal disruption	Using synthe-sized Kin voices	Proactive speaker
Methodology	TTS API, extractive Q&A, abstractive summarization	Contrastive sentiment analysis, Natural Language Inference, Media consumption. Serendipity in news coverage	ACS-QG	SSML to include prosody modification	QAGen, distractor Generator (DG)	speech synthesis API of Naver Clova	QP graph	Google wavenet TTS	Reco style is varied	Webspeech API and google Speech synthesis	Sensors to monitor the users	kinVoice- Alexa based VUI.	Using sensors and voice interface
Advantages	Different voices to Q & A makes it less monotonous	Media consumption. Serendipity in news coverage	Generation of questions are controlled,	Prosody modification helps in emphasizing the key answers	New dataset for quiz style	Interactions helps retain news	Avoid repetition to retain consumer retention	Quick access to news	Personalised recommendation	Short response makes efficient interaction	Opportune movements found to retain attention	Bridge physical distance	Starter conversation should be human-like
Drawbacks	Imperfect speech & Q&A, lack of editorial judgment	Referring to too many articles could be misleading		Unnatural voice due to modification	grammatical errors were present	participants not ready to ask questions in between	unable to handle small talk, QA is not accurate	Participants could identify TTS, can it be more human-like	less topics to discuss	not customizable	Camera was used to monitor the participants. Privacy concern	Synthetic kinVoice not developed well	experiment done in dormitory.
Evaluations:	user studies and baseline analysis	None	Quantitative and Qualitative	User study with TTS vs Prosody intervention	Case study: google platform with auto QA	Wizard of ox style evaluation	engage with multi-turn questions	3 tests with human voice, TTS and hybrid	Alexa users, with different types of engagement	3 response styles	a week-long field study on 40 participants	ser study on 4 cases with kin-Voice,	field study 1 week
Results:	Less monotonous but diverse questions	Too many resources are misleading	Quality of QG outperforms baseline	Lower speaking rate and increasing pitch points to answer	Users preferred these Quizzes	Participants not ready to ask questions in between	People continued to have longer conversation	Acceptance of integration of voices	Higher acceptance rate for entity based recommendation	Short responses were preferred, but not for news	Privacy concern	For daily tasks preferred kinVoice, others eerie.	Utterance interruption was preferred.
Voice Interactivity	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
news	Yes	No	Yes	No	No	Yes	Yes	No	Yes	No	No	No	No

A. Components for designing a Vocal assistant for reading news

From this literature review, we propose the components needed for designing a vocal assistant for reading news.

Interface: First, we need to consider the device where the vocal assistant for reading news will be included. It could be implemented in Smartphones, smart speakers, or smart displays. Screens are beneficial to highlight key points, however, Chuklin et al in [5], propose using prosody modification to highlight key answers via audio modulation. The choice of whether or not to use a screen could be left up to the user.

Speech recognition: When the user converses with the interface, the next step is to understand the query. So Speech recognition needs to be done. From the query, the intent of the query needs to be extracted. Natural Language Processing is done on the query by using various algorithms like entity recognition, and sentiment analysis.

Content extraction: From the intent of the query, reputable sources need to be extracted from the internet. While obtaining resources from the internet, the preference of the user should be taken into account to provide personalized recommendation. Summarization techniques are applied to the resources to reduce the burden on the user by presenting shorter versions of lengthy documents after obtaining them. To make the news sound more engaging, QA generators can be used to form Q&A sessions between two voices as suggested by Laban et al in [12]. In order to avoid Filter Bubble effect, contrastive news can also be provided to encourage users' curiosity and provide serendipity. To measure if the users are attentive during news briefings, quizzes can be curated to make the experience educational and enjoyable as used by Lelkes et al in [14].

Voice and tone: Designing the voice and tone of the assistant to be clear, neutral and easy to understand. With advancements in Text-To-Speech, and SSML we can modify the speaking rate, pitch, and tone to make the voice natural. As proposed in [3], the synthesised voice of friends can be used to feel comfortable discussing news with the assistant.

Interactivity: While designing voice assistant, interactivity between the user and assistant should be considered. [10], [12], [19], are models that suggest interactivity is important to retain user's attention. Hence, while obtaining resources to present to the user, the voice assistant should be able to clarify any doubts when the user's query has multiple intents. It should allow users to interrupt, and propose breaks during topic changes to ask any questions they may have. Interactivity can be achieved by making the virtual assistant Proactive by looking at opportune moments to converse with the user.

The final component to consider while designing any system is: Privacy and security: Privacy is a major concern for users, as the constant listening feature of these devices can lead to fears of surveillance. To address this, users should be well-informed about the device's privacy settings and the device should have a method for disabling listening in order to maintain privacy. User data like historical conversation logs should be securely stored and not shared with third parties

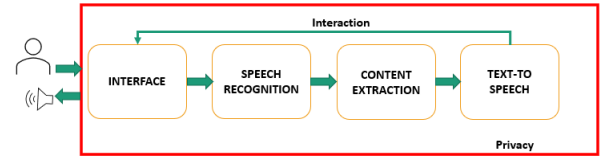


Fig. 9. Proposed Architecture

without user consent. These data can be used to personalize the recommendation for the users. There should also be a limit of logs that can be stored and an option to delete all the data should be made available.

In Fig 9, we propose an architecture to design Vocal assistants by following the guidelines presented above.

VI. CONCLUSION

This study explores ways to enhance the interaction between vocal conversational agents and users while reading news. Existing voice assistants have limitations, such as repetitive responses and unnatural speech, which erode user trust. Our research review focuses on recent advancements in improving voice assistant interaction, including improvements in QAGen to reduce repetition and advances in Text-To-Speech technology for more natural speech. We also found that automatically generated Podcasts can increase news consumption and allow interaction during news reading creating a human-like conversation experience. Personalized news recommendations increase user engagement, though diversity may be lacking. We present guidelines to consider while designing these models and maintain the privacy of the users.

This literature review is not an extensive study as it does not include all state-of-the-art models. However, it can be used as a guideline for those developing voice assistants. As a possible future work, the authors propose to implement a vocal assistant for reading news, following the guidelines proposed in this study.

REFERENCES

- [1] Marcelo Carvalho Afonso, Pedro Almeida, Pedro Beça, Telmo Silva, and Iulia Covalenco. Usability of text-to-speech technology in creating news podcasts using portuguese of portugal. In *ACM International Conference on Interactive Media Experiences*, pages 363–368, 2022.
- [2] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. Hello there! is now a good time to talk? opportune moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–28, 2020.
- [3] Sam WT Chan, Tamil Selvan Gunasekaran, Yun Suen Pai, Haimo Zhang, and Suranga Nanayakkara. Kinvoices: Using voices of friends and family in voice interfaces. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.
- [4] Li Chen, Zhirun Zhang, Xinzhi Zhang, and Lehong Zhao. A pilot study for understanding users' attitudes towards a conversational agent for news recommendation. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6, 2022.
- [5] Aleksandr Chuklin, Aliaksei Severyn, Johanne Trippas, Enrique Alfonso, Hanna Silen, and Damiano Spina. Prosody modifications for question-answering in voice-only settings. *arXiv preprint arXiv:1806.03957*, 2018.

- [6] Henry Kudzanai Dambanemuya and Nicholas Diakopoulos. Auditing the information quality of news-related queries on the alexa voice assistant. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.
- [7] Mateusz Dubiel, Alessandra Cervone, and Giuseppe Riccardi. Inquisitive mind: a conversational news companion. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3, 2019.
- [8] Anushka Gupta, Diksha Chugh, Rahul Katarya, et al. Automated news summarization using transformers. In *Sustainable Advanced Computing*, pages 249–259. Springer, 2022.
- [9] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. Keep it short: A comparison of voice assistants’ response behavior. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2022.
- [10] Hyunhoon Jung, Changhoon Oh, Gilhwan Hwang, Cindy Yoonjung Oh, Joonhwan Lee, and Bongwon Suh. Tell me more: Understanding user interaction of smart speaker news powered by conversational search. In *Extended abstracts of the 2019 chi conference on human factors in computing systems*, pages 1–6, 2019.
- [11] Philippe Laban, John Canny, and Marti A Hearst. What’s the latest? a question-driven news chatbot. *arXiv preprint arXiv:2105.05392*, 2021.
- [12] Philippe Laban, Elicia Ye, Srulay Korlakunta, John Canny, and Marti Hearst. Newspod: Automatic and interactive news podcasts. In *27th International Conference on Intelligent User Interfaces*, pages 691–706, 2022.
- [13] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–31, 2018.
- [14] Adam D Lelkes, Vinh Q Tran, and Cong Yu. Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021*, pages 2501–2511, 2021.
- [15] Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043, 2020.
- [16] Sara Irma Parnell, Stefanie Helene Klein, and Franziska Gaiser. Do we know and do we care? algorithms and attitude towards conversational user interfaces: Comparing chatbots and voice assistants. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6, 2022.
- [17] Shruti Rao, Valeria Resendez, Abdallah El Ali, and Pablo Cesar. Ethical self-disclosing voice user interfaces for delivery of news. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–4, 2022.
- [18] Harshita Sahijwani, Jason Ingyu Choi, and Eugene Agichtein. Would you like to hear the news? investigating voice-based suggestions for conversational news recommendation. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 437–441, 2020.
- [19] Jing Wei, Tilman Dingler, and Vassilis Kostakos. Understanding user perceptions of proactive smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–28, 2021.